Short Communication

Genetic Variation and Willingness to Participate in Epidemiologic Research: Data from Three Studies

Parveen Bhatti,[1] Alice J. Sigurdson,[1] Sophia S. Wang,[2] Jinbo Chen,[3] Nathaniel Rothman,[4] Patricia Hartge,[5] Andrew W. Bergen,[6] and Maria Teresa Landi[6]

[1]Radiation Epidemiology Branch, [2]Hormonal and Reproductive Epidemiology Branch, [3]Biostatistics Branch, [4]Occupational and Environmental Epidemiology Branch, [5]Epidemiology and Biostatistics Program, and [6]Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Department of Health and Human Services, Bethesda, Maryland

Abstract

The differences in common genetic polymorphism frequencies by willingness to participate in epidemiologic studies are unexplored, but the same threats to internal validity operate as for studies with nongenetic information. We analyzed single nucleotide polymorphism genotypes, haplotypes, and short tandem repeats among control groups from three studies with different recruitment designs that included early, late, and never questionnaire responders, one or more participation incentives, and blood or buccal DNA collection. Among 2,955 individuals, we compared 108 genotypes, 8 haplotypes, and 9 to 15 short tandem repeats by respondent type. Among our main comparisons, single nucleotide polymorphism genotype frequencies differed significantly ($P < 0.05$) between respondent groups in six instances, with 13 expected by chance alone. When comparing the odds of carrying a variant among the various response groups, 19 odds ratios were ≤ 0.70 or ≥ 1.40, levels that might be notably different. Among the various respondent group comparisons, haplotype and short tandem repeat frequencies were not significantly different by willingness to participate. We observed little evidence to suggest that genotype differences underlie response characteristics in molecular epidemiologic studies, but a greater variety of genes should be examined, including those related to behavioral traits potentially associated with willingness to participate. To the extent possible, investigators should evaluate their own genetic data for bias in response categories. (Cancer Epidemiol Biomarkers Prev 2005;14(10):2449–53)

Introduction

Loss of information because of nonresponse can compromise the validity of risk estimates from epidemiologic studies, which is a growing concern in light of declining participation rates (1-3). For various behaviors, exposures, and outcomes, numerous studies have investigated the potential effects of nonresponse (2, 4-9), but corresponding threats due to genetic variation are unexplored; validity in genetic studies is not assured because we assume a genetic variant is unrelated to response (3).

Although genetic variation with "true" nonresponse (i.e., those who did not provide genetic material) is impossible to address, genetic studies with recruitment waves provide a unique opportunity to investigate genetic frequency differences by participation. We examined frequencies of single nucleotide polymorphism genotypes, haplotypes, and short tandem repeat alleles by response status in control subjects from three studies with different recruitment designs allowing comparisons of early, late, and never questionnaire responders, one or more participation incentives, and blood or buccal DNA donation.

Materials and Methods

Subjects. Study A participants were controls in a nested-case control study of breast cancer among the U.S. Radiologic Technologists cohort (10, 11). All controls were female cohort members that provided consent and a blood sample for genetic analyses and for whom a study survey had been previously mailed. Among them were early responders ($n = 679$), late responders, requiring an extra incentive (a one dollar bill) to participate ($n = 54$), and nonresponders ($n = 50$) to the previously mailed questionnaire. Because sampling for the breast cancer case-control study occurred independently of questionnaire response, nonrespondents were included in the biospecimen recruitment effort.

Study B participants consisted of non-Hispanic Caucasian controls (516 males and 466 females) recruited for a case-control study of non-Hodgkin's lymphoma, from within four areas of the Surveillance, Epidemiology, and End Results cancer registry of the National Cancer Institute (12). Of these, 554 controls chose to provide a blood sample for genetic analyses, whereas 209 controls who did not provide blood samples did provide saliva (buccal) samples; 741 of these subjects responded to biospecimen donation at the time of study questionnaire administration (regular responders), whereas 22 subjects who initially refused to provide blood or buccal cell samples provided buccal cells after a final mail query at the end of the study (late responders).

Study C participants were controls that provided blood samples (232 females and 958 males) from a case-control study of lung cancer from the Lombardy region of Italy. Two-hundred and fifty-two controls (less incentivized group) were recruited by mail and telephone follow-up; the invitation was accompanied by cash or gas coupons and by a letter endorsing the study signed by the subjects' family physician. Nine-hundred and thirty-eight controls (highly incentivized group) were recruited using a letter of invitation, accompanied by a direct call by the subjects' family physician, a letter

from the mayor of the participating cities supporting the research, and gas coupons to the subjects and family physicians; a toll-free number through which potential participants could obtain information about the study was also established and television advertisements were made.

Laboratory Methods. Study A participants were genotyped for 36 single nucleotide polymorphisms in DNA repair and growth factor genes (13). All samples from studies B and C were analyzed at the Core Genotyping Facility of the National Cancer Institute (http://cgf.nci.nih.gov/home.cfm). Study B participants were genotyped for 103 single nucleotide polymorphisms in genes involved in immune, oxidative stress, metabolism, cell cycle, and DNA repair pathways. For short tandem repeat analysis in all three studies, samples were quantified using PicoGreen and reverse transcription-PCR analysis and profiled using the Applied Biosystems Identifiler kit. Fifteen short tandem repeat loci were analyzed in studies A and C, and nine were analyzed in study B.

Statistical Analysis. We only considered those single nucleotide polymorphisms with a minor allele frequency of ≥5% for analyses; 15 single nucleotide polymorphisms from study A and 16 single nucleotide polymorphisms from study B were too infrequent for inclusion. We reconstructed haplotypes for *APEX, BACH1, BRCA2, TGFβ1, XRCC1,* and *ZNF350* for study A and for *IL10* and *LTA/TNF* for study B (separately for each comparison group) using the PHASE software package (14). Haplotypes were not reconstructed for regular versus late responder analyses in study B because of the small number of late responders. We analyzed single nucleotide polymorphism and haplotype frequencies among categories of study recruitment using contingency table analyses in SAS release 8.02 (SAS Institute, Inc., Cary, NC); in addition to χ^2 analyses and odds ratios comparing the frequency of single nucleotide polymorphism carriers and noncarriers between the various comparison groups, single nucleotide polymorphism genotypes (homozygous wild type, heterozygous, homozygous variant)

were analyzed among participation categories using the Mantel-Haenszel test for trend. We noted odds ratios that were ≤0.70 or ≥1.40 because this magnitude is approximately symmetrical around 1.0, and values outside this range could conceivably impact genotype-disease associations. For short tandem repeat analysis, we used SAS release 8.02 to calculate short tandem repeat genotype means and ranges at each locus for the various comparison groups. In addition, using Arlequin version 2000 (15), we estimated the standardized fixation index or F_{ST} (ratio of the number of different alleles observed between two individuals in two different samples compared with the number of different alleles observed between two individuals in the same sample; ref. 16). F_{ST} provides a single measure of genetic differentiation when multiallelic loci are being considered, such as short tandem repeats. All tests for significance were two-sided with α set at 0.05.

Results

When comparing late responders to early responders and nonresponders to early responders in study A (Table 1), we found that seven and eight odds ratios were ≤0.70 or ≥1.40, respectively. The *TGFβ1* P25R variant differed significantly (trend test, $P = 0.03$) among the questionnaire response groups; one statistically significant trend was expected by chance. Haplotype frequencies for the various genes were not found to be statistically significantly different (χ^2 test; not shown).

Two of the variant frequencies were significantly different between the blood and buccal groups in study B; these were *EPHX1* H139R ($P = 0.0064$) and *CYP1B1* V432L ($P = 0.027$; Fig. 1); at least four were expected to differ by chance. The Mantel-Haenszel test for trend revealed significant frequency differences between the biospecimen groups with increasing copies of *EPHX1* H139R ($P = 0.0031$). Significant trends were also found with *IL8RB* 1235T>C ($P = 0.024$) and *MPO* 642G>A ($P = 0.011$). Four of the 87 single nucleotide polymorphisms in

Table 1. Odds ratios for late and never responders compared with early responders to a mailed questionnaire in study A

Gene	Polymorphism	Late responders ($n = 54$), odds ratio*	Nonresponders ($n = 50$), odds ratio†	Pearson χ^2 (P)‡	χ^2 trend (P)§
DNA repair					
ATM	P1526P‖ (rs1800889)	2.56	3.70	0.17	0.07
APEX	D148E (rs1130409)	0.89	1.42	0.45	0.73
APEX	Q51H (rs1048945)	2.27	0.79	0.45	0.90
BRCA2	N372H (rs2421655)	0.88	1.47	0.37	0.43
BRCA2	N289H (rs766173)	1.64	1.11	0.71	0.61
BRCA2	T1915M (rs4987117)	1.11	0.77	0.87	0.72
BACH1	−64G>A¶ (rs2048718)	0.95	0.78	0.74	0.74
BACH1	P919S (rs4986764)	0.93	0.86	0.88	0.27
POLD1	R119H (rs1726801)	1.16	2.04	0.37	0.17
XRCC1	R194W (rs1799782)	0.87	0.93	0.94	0.91
XRCC1	R399Q (rs25478)	0.56	1.06	0.17	0.55
XRCC1	R280H (rs25489)	2.27	1.01	0.51	0.60
ZNF350	L66P (rs2278420)	0.91	1.00	0.95	0.70
ZNF350	R501S (rs2278415)	0.79	1.27	0.59	0.65
ZNF350	S472P (rs4986771)	0.74	0.53	0.41	0.41
ZNF350	1845C>T (rs4986770)	0.77	0.93	0.78	0.64
Growth factors					
ERBB2	I655V (rs1801200)	0.80	0.61	0.34	0.36
IGFBP2	202A>C (rs2854744)	0.79	1.33	0.55	0.77
TGFB1	L10P (rs1982073)	1.04	0.93	0.95	0.44
TGFB1	T263I (rs1800472)	0.70	1.40	0.72	0.90
TGFB1	P25R (rs1800471)	0.63	0.43	0.08	0.03

*Wild-type late responders versus wild-type early responders ($n = 679$); reflects frequency of genotyping successes and does not include individual failures.
†Wild-type nonresponders versus wild-type early responders ($n = 679$); reflects frequency of genotyping successes and does not include individual failures.
‡Pearson χ^2 test for single nucleotide polymorphism alleles (variant carriers versus noncarriers among the response groups).
§Test for trend for single nucleotide polymorphism genotypes (homozygote wild type, heterozygote, and homozygote variant).
‖Amino acids and their symbols: H, histidine; R, arginine; Q, glutamine; A, alanine; S, serine; F, phenylalanine; P, proline; K, lysine; L, leucine; N, asparagine; I, isoleucine; D, aspartic acid; GL, glycine; Y, tyrosine; C, cytosine; V, valine; E, glutamic acid; T, threonine (Entrez SNP reference single nucleotide polymorphism number in parentheses).
¶Nucleotides and their symbols: A, adenine; T, thymine; G, guanine; C, cytosine.
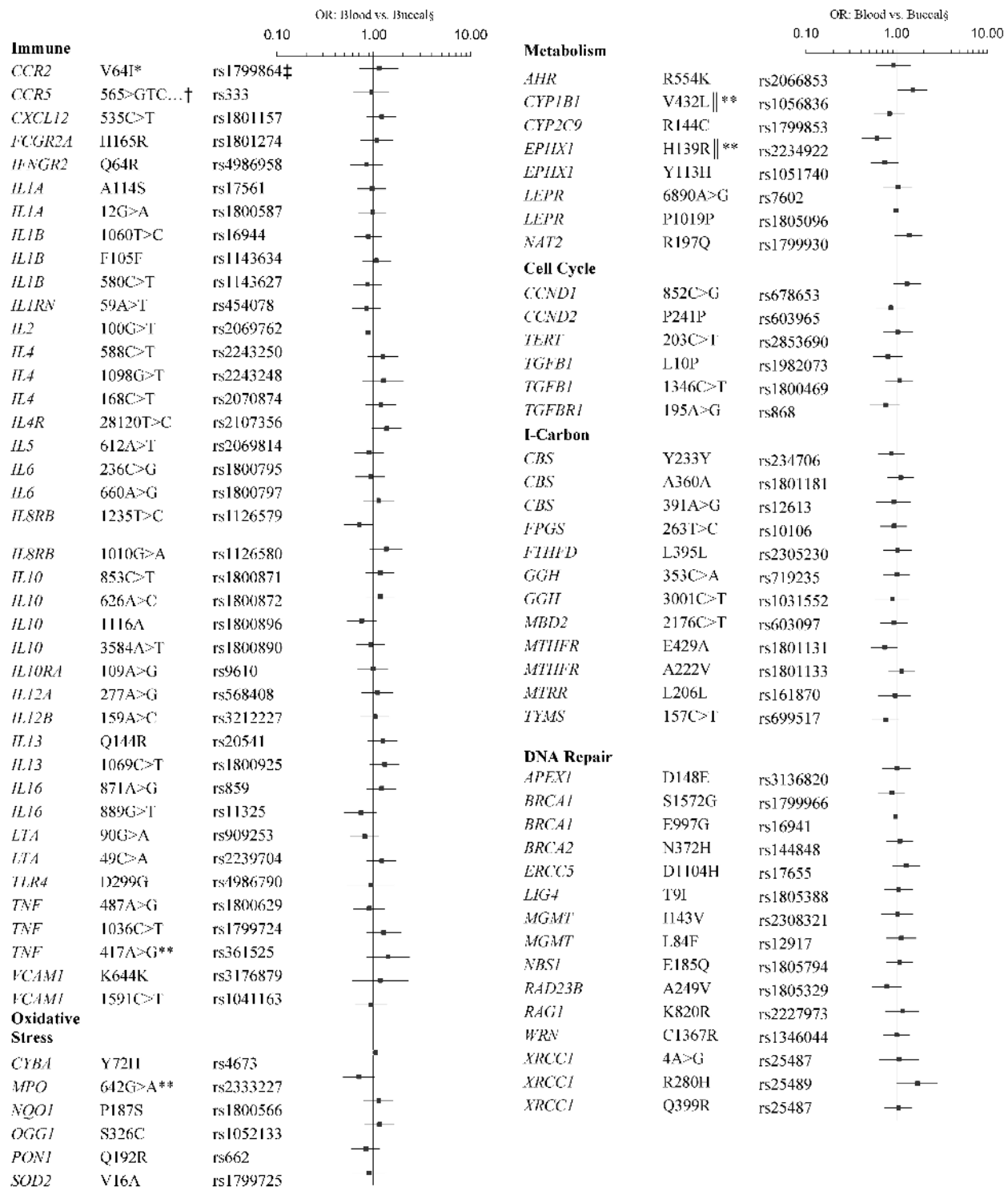
Figure 1. Study B single nucleotide polymorphism frequency comparison (blood versus buccal sample participants).

* Amino acids and their symbols: H: Histidine, R: Arginine, Q: Glutamine, A: Alanine, S: Serine, F: Phenylalanine, P: Proline, K: Lysine, L: Leucine, N: Asparagine, I: Isoleucine, D: Aspartic Acid, GL: Glycine, Y: Tyrosine, C: Cytosine, V: Valine, E: Glutamic Acid, T: Threonine;
† Nucleotides and their symbols: A: Adenine, T: Thymine, C: Cytosine, G: Guanine
‡ Entrez SNP reference SNP ID number
§ wildtype carriers among blood sample providers compared to buccal sample providers
‖ $p < 0.05$ (Pearson chi-square carriers vs. non-carriers)
** OR ≤ 0.70 or ≥ 1.40

Fig. 1 had point estimates that were ≤ 0.70 or ≥ 1.40. Haplotype frequencies were not found to be significantly different between the biospecimen groups (interleukin-10, $P = 0.25$; LTA/TNF, $P = 0.45$). Among the respondent groups, six single nucleotide polymorphism frequencies were significantly different: *IL1A* A114S, *IL1A* 12G>A, *IL4R* 28120T>C, *NQO1* P187S, *TYMS* 157C>T, and *MGMT* L84F (not shown); eight statistically significant differences were expected by chance.

Table 2.

A. Short tandem repeat among early vs late and early vs nonresponders

Short tandem repeat	Early responders (*n* = 679)	Late responders (*n* = 54)			Nonresponders (*n* = 50)		
	Mean (range)	Mean (range)	F_{ST}*	*P*	Mean (range)	F_{ST}*	*P*
CSF1PO	11.2 (7-16)	11.3 (9-14)	−0.0039	0.87	11.2 (9-15)	0.0051	0.12
D13S317	11.0 (5-15)	11.0 (8-14)	−0.0006	0.47	11.4 (8-14)	0.0060	0.06
D16S539	11.4 (8-15)	11.4 (8-14)	−0.0015	0.58	11.5 (8-14)	−0.0037	0.87
D18S51	15.0 (9-25)	14.7 (10-20)	−0.0005	0.48	14.9 (11-22)	0.0000	0.45
D19S433	14.0 (11-18.2)	14.0 (12-16.2)	−0.0035	0.92	14.0 (12-16.2)	−0.0015	0.51
D21S11	30.0 (25-34.2)	29.7 (27-33.2)	0.0053	0.05†	30.0 (27-34.2)	0.0048	0.07
D2S1338	20.4 (15-28)	20.3 (16-26)	0.0041	0.07	20.4 (16-26)	−0.0005	0.50
D3S1358	16.1 (11-20)	16.0 (11-19)	−0.0039	0.94	16.1 (14-20)	−0.0041	0.93
D5S818	11.6 (7-15)	11.4 (7-14)	0.0084	0.06	11.7 (9-14)	0.0026	0.21
D7S820	10.1 (7-14)	10.3 (7-14)	−0.0039	0.97	10.1 (7-14)	−0.0012	0.54
D8S1179	12.9 (8-17)	12.6 (8-16)	0.0046	0.08	13.0 (8-17)	−0.0022	0.72
FGA	22.2 (17-28)	22.6 (18-27)	0.0019	0.24	22.5 (18-27)	0.0053	0.05
TH01	8.0 (5-10)	7.6 (5-10)	0.0125	0.01	8.1 (6-9.3)	0.0001	0.41
TPOX	9.1 (8-12)	9.0 (8-12)	−0.0022	0.62	9.2 (8-12)	−0.0025	0.61
vWA	16.7 (13-21)	16.6 (13-21)	−0.0030	0.83	16.6 (14-20)	−0.0028	0.80

B. Study B short tandem repeats among blood vs buccal sample participants

Short tandem repeat	Blood (*n* = 554)	Buccal (*n* = 209)	F_{ST}	*P*
	Mean (range)	Mean (range)		
D13S317	11 (8-15)	11.1 (8-15)	0.00014	0.36
D18S51	15 (9-23)	14.7 (0-26)	−0.00082	0.68
D21S11	29.8 (24.2-34.2)	29.8 (26-35.2)	−0.00038	0.56
D3S1358	16.1 (11-20)	16.1 (13-19)	−0.00024	0.49
D5S818	11.6 (7-15)	11.6 (9-15)	−0.00062	0.65
D7S820	10.1 (7-14)	10 (7-15)	−0.00075	0.83
D8S1179	12.8 (8-17)	12.8 (8-17)	−0.00083	0.84
FGA	22.1 (17-28)	21.6 (0-27)	−0.00019	0.52
vWA	16.8 (13-21)	16.7 (13-19)	0.00037	0.31

C. Study C short tandem repeat analysis among less vs highly incentivized responders

Short tandem repeat	Less incentivized (*n* = 252)	Highly incentivized (*n* = 938)	F_{ST}	*P*
	Mean (range)	Mean (range)		
CSF1PO	11.2 (7-15)	11.1 (0-15)	0.0017	0.07
D13S317	11.1 (8-15)	11.0 (7-15)	0.00065	0.18
D16S539	11.3 (8-15)	11.3 (8-15)	−0.00096	0.95
D18S51	14.9 (10-24)	14.8 (0-24)	0.00048	0.19
D19S433	13.9 (10-18)	13.9 (10-18)	−0.00052	0.69
D21S11	30.0 (24-35)	29.9 (24-35)	−0.00036	0.62
D2S1338	20.0 (15-27)	20.1 (14-27)	0.00072	0.15
D3S1358	16.2 (12-19)	16.1 (10-19)	0.00023	0.32
D5S818	11.5 (8-14)	11.6 (8-15)	−0.00013	0.43
D7S820	10.2 (7-14)	10.1 (0-14)	−0.00058	0.71
D8S1179	13.0 (8-17)	12.9 (8-17)	−0.00014	0.49
FGA	22.2 (17-28)	22.2 (0-29)	0.00014	0.34
TH01	7.8 (6-13)	7.9 (4-10)	0.00032	0.26
TPOX	9.1 (5-12)	9.2 (5-13)	0.00018	0.32
vWA	16.6 (13-20)	16.7 (12-21)	−0.00071	0.81

*Compared with early responders.
†*P* = 0.0488.

Table 2 shows short tandem repeat results for all three studies. In study A, we found two loci (*D21S11*, *TH01*) that were statistically significantly different between early and late responders; one difference was expected by chance. We found no statistically significant F_{ST} values in studies B and C; we also found no statistically significant F_{ST} values when considering the early and late respondent groups in study B (not shown).

Discussion

To our knowledge, this is the first exploration of the threat to internal validity from genotype frequency differences by participation status for cancer genetic epidemiology. In the present analysis, we did not find that genotype frequency differences between categories of respondents and incentive groups significantly exceeded the number expected by chance. The biases that occur in epidemiologic studies of the effects of genetic variants correspond to the general framework for any exposure (17). That is, biases may be related to inclusion in the study (selection bias), to availability or accuracy of response (recall or ascertainment bias), and to correlation with other factors (confounding, model misspecification, population stratification). Although commentators have recently focused much attention on population stratification (18-21), a form of confounding, selection bias and response bias have had less attention, in part,

because genetic data on nonresponders are difficult by definition to obtain.

Because polymorphism frequencies in nonresponders are unknown, investigators have assumed that participation in genetic studies was unrelated to genotype. This may not be true when variants in genes related to behavioral characteristics are under investigation or if a variant may be related to family disease history; willingness to participate has been associated with family history of the particular disease under study (9). In our analyses, there were a few statistically significant differences by participation status; whereas the number of such observations was consistent with expectation, there were no statistically significant differences consistent within and between studies. We also found no evidence of differences, beyond those expected by chance, between subjects opting to provide mouthwash samples for genetic analysis instead of blood samples.

As with confounding, a statistically significant association between a genetic variant and response is not necessary for bias to occur; a sufficient relationship must simply exist in the data (3). Thus, we have identified those single nucleotide polymorphisms in Table 1 and Fig. 1 with response differentials (odds ratios ≤0.70 or ≥1.40) that may result in substantial bias under certain circumstances; although not examined in these series of analyses, for biased odds ratios to occur in case-control studies, response differentials must themselves be different between cases and controls. The mathematics of participation bias has been described elsewhere (22).

The present analysis had several strengths in that multiple studies with polymorphisms in common permitted exploration and confirmation of study specific findings and each study provided data on plausible surrogates for nonresponse, such as reaction to incentives. Study A, in particular, provided a rare opportunity to assess genetic profiles of questionnaire nonresponders. A limitation was that the polymorphisms we examined were already available in these three studies; they were selected based on *a priori* disease associations, not as candidate variants in genes potentially related to willingness to participate.

Despite the apparent conundrum of assessing genetic characteristics of "true" nonresponders, we show there are opportunities to approach the question of response bias in molecular epidemiologic studies. Our findings, while reassuring, cannot exclude that differences by response exist in other genes. The potential for bias due to the "genetics of response" should continue to be evaluated, when possible, within the wider molecular epidemiologic research community.

Acknowledgments

References

1. Austin MA, Criqui MH, Barrett-Connor E, Holdbrook MJ. The effect of response bias on the odds ratio. Am J Epidemiol 1981;114:137–43.
2. Criqui MH, Barrett-Connor E, Austin M. Differences between respondents and non-respondents in a population-based cardiovascular disease study. Am J Epidemiol 1978;108:367–72.
3. Hartge P. Raising response rates: getting to yes. Epidemiology 1999;10: 105–7.
4. Chen R, Wei L, Syme PD. Comparison of early and delayed respondents to a postal health survey: a questionnaire study of personality traits and neuropsychological symptoms. Eur J Epidemiol 2003;18:195–202.
5. Holt VL, Daling JR, Stergachis A, Voigt LF, Weiss NS. Results and effect of refusal recontact in a case-control study of ectopic pregnancy. Epidemiology 1991;2:375–9.
6. Jackson R, Chambless LE, Yang K, et al. Differences between respondents and nonrespondents in a multicenter community-based study vary by gender ethnicity. The Atherosclerosis Risk in Communities (ARIC) Study Investigators. J Clin Epidemiol 1996;49:1441–6.
7. Paganini-Hill A, Hsu G, Chao A, Ross RK. Comparison of early and late respondents to a postal health survey questionnaire. Epidemiology 1993;4: 375–9.
8. Voigt LF, Koepsell TD, Daling JR. Characteristics of telephone survey respondents according to willingness to participate. Am J Epidemiol 2003; 157:66–73.
9. Wang SS, Fridinger F, Sheedy KM, Khoury MJ. Public attitudes regarding the donation and storage of blood specimens for genetic research. Community Genet 2001;4:18–26.
10. Doody MM, Sigurdson AS, Kampa D, et al. Randomized trial of financial incentives and delivery methods for improving response to a mailed questionnaire. Am J Epidemiol 2003;157:643–51.
11. Sigurdson AJ, Doody MM, Rao RS, et al. Cancer incidence in the US radiologic technologists health study, 1983-1998. Cancer 2003;97:3080–9.
12. Chatterjee N, Hartge P, Cerhan JR, et al. Risk of non-Hodgkin's lymphoma and family history of lymphatic, hematologic, and other cancers. Cancer Epidemiol Biomarkers Prev 2004;13:1415–21.
13. Sigurdson AJ, Hauptmann M, Chatterjee N, et al. Kin-cohort estimates for familial breast cancer risk in relation to variants in DNA base excision repair, BRCA1 interacting and growth factor genes. BMC Cancer 2004;4:9.
14. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–89.
15. Arlequin ver. 2.000: A software for population genetics data analysis. 2000.
16. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 1992;131:479–91.
17. Rothman K, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.
18. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11:505–12.
19. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 2000;92:1151–8.
20. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002;11:513–20.
21. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in case-control association studies of admixed populations. Genet Epidemiol 2004;27:14–20.
22. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont (CA): Lifetime Learning Publications; 1982.